Data Security and Privacy

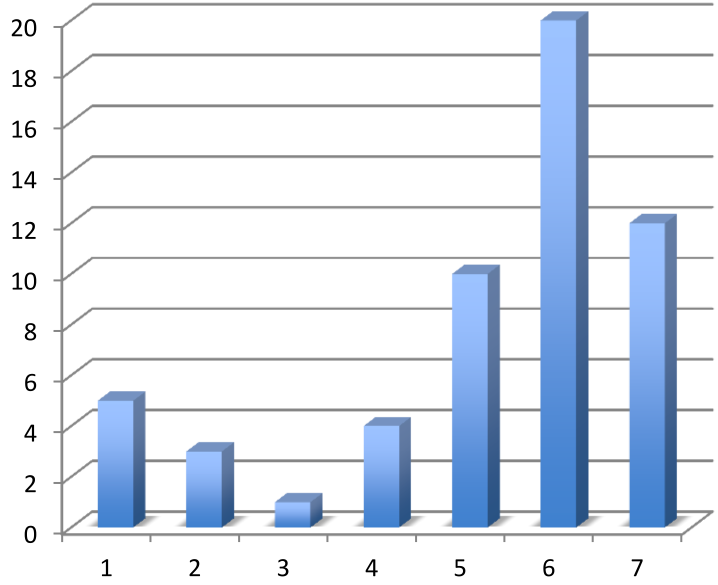
Topic 21: Publishing Private Histogram and Using it for Classification

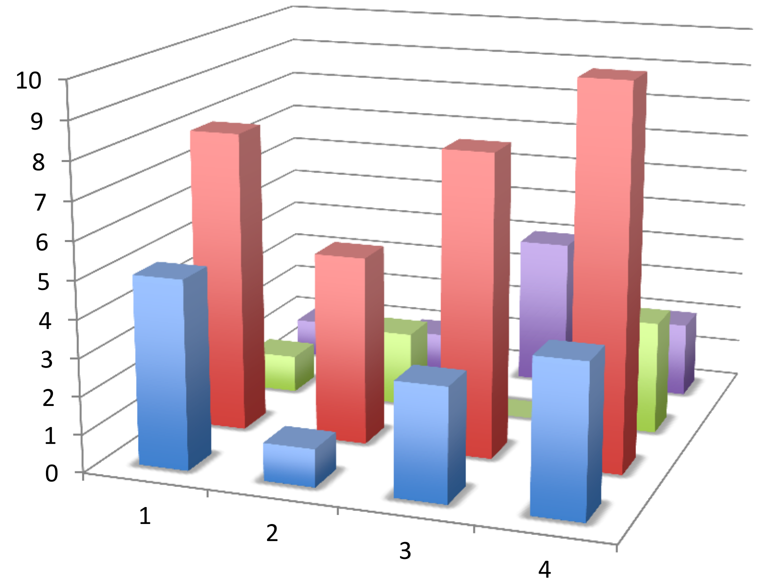
Reading

- Wahbeh H. Qardaji, Weining Yang, Ninghui Li: Differentially private grids for geospatial data. ICDE 2013: 757-768
- Dong Su, Jianneng Cao, Ninghui Li, Min Lyu: PrivPfC: differentially private data publication for classification. VLDB J. 27(2): 201-223 (2018)

Histogram

- A histogram is a graphical representation of the distribution of numerical data





Noisy Histograms

- A histogram is a graphical representation of the distribution of numerical data
  - a partitioning of the data domain into multiple non-overlapping bins
  - the number of data points in each bin

- By adding suitable noises, publishing histogram satisfies DP

Using Histogram to Answer Range Queries

- A range query represents a hyperrectangle in the d-dimensional domain specified by the dataset, and asks for the number of tuples that fall within the bins that are completely included in the area covered by the hyperrectangle

2	3	4	1	5	1	0	6

5	**1**	**3**	**4**
8	5	8	10
1	2	0	3
1	1	4	2

Utility Metrics for Range Queries (1)

- Mean Absolute Error (MAE)
  - absolute difference between the noisy answer and the true answer
- Mean Squared Absolute Error (MSAE)
  - often easier to compute
  - MSAE is the variance of the random noise

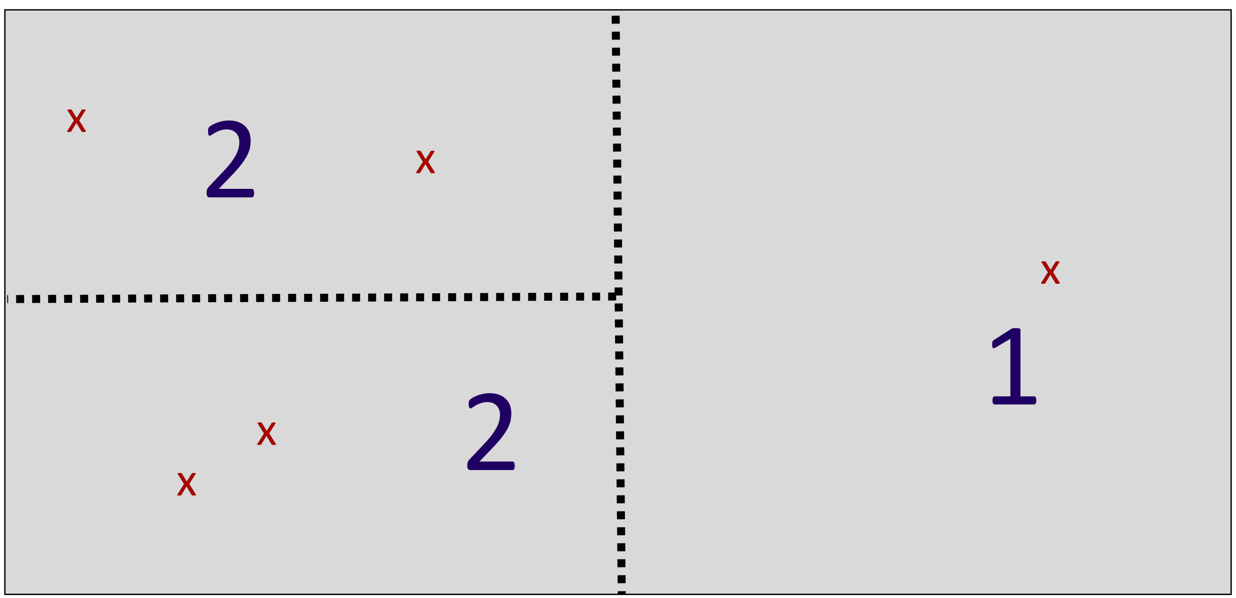
Utility Metrics for Range Queries (2)

- Mean Relative Error (MRE)
  - impact of the same absolute error is different when the true answers are different
  - the true answer may be very small, or even 0
    - chooses a threshold θ to be used as the denominator

$$\text{relative error} = \frac{|\text{true anser} - \text{obtained answer}|}{\max(\theta, \text{true anser})}$$

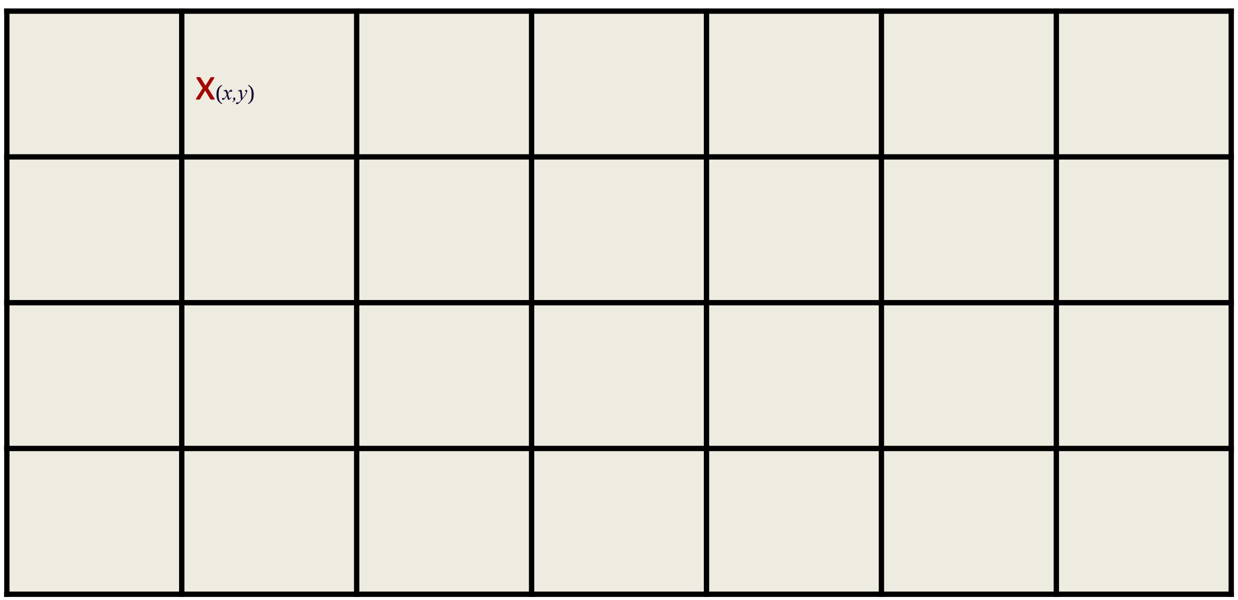
Example: Geospatial Data

Example: Geospatial Data



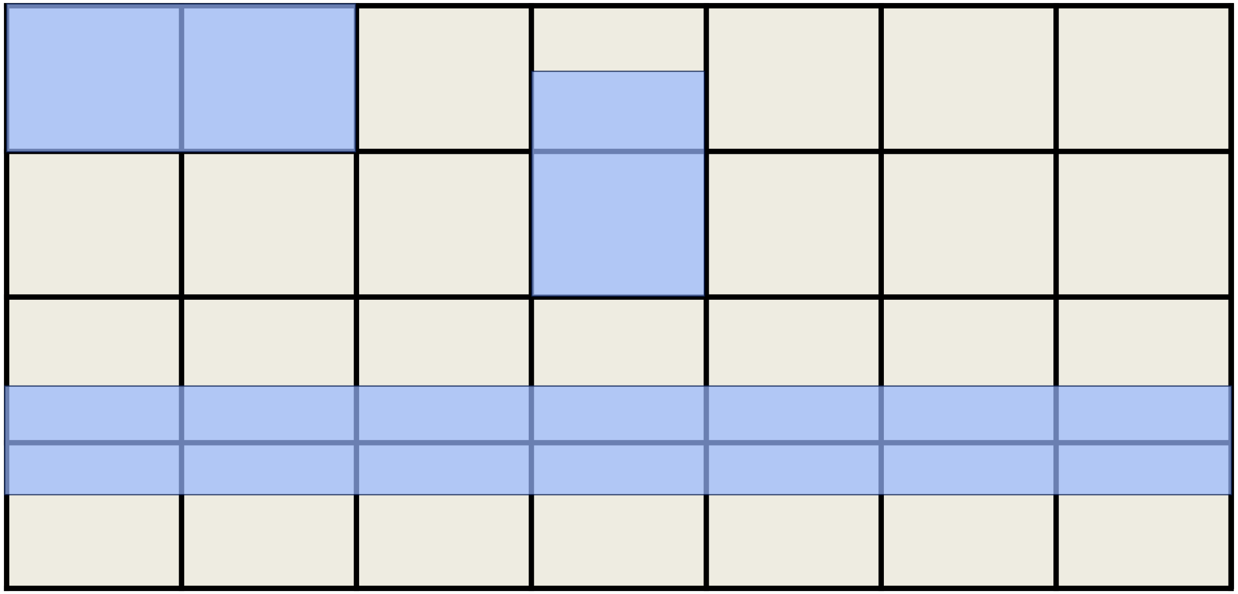
Uniform Grid

- Partition domain into m x m cells of equal size
- Add noise to counts of each cell to satisfy differential privacy



Measuring Utility

- Error from answering range queries
  - a query is a rectangle in the data domain



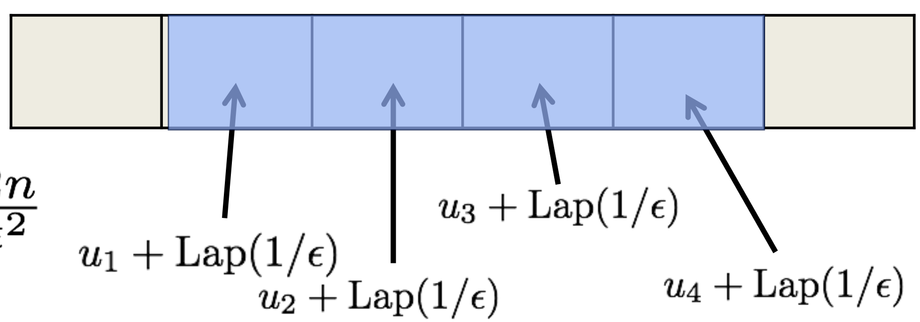
Sources of Error

1. Error from satisfying Differential Privacy (noise error)

- Adding noise from the Laplace Distribution

$$\text{Var}\left(\text{Lap}(1/\epsilon)\right) = \frac{2}{\epsilon^2}$$

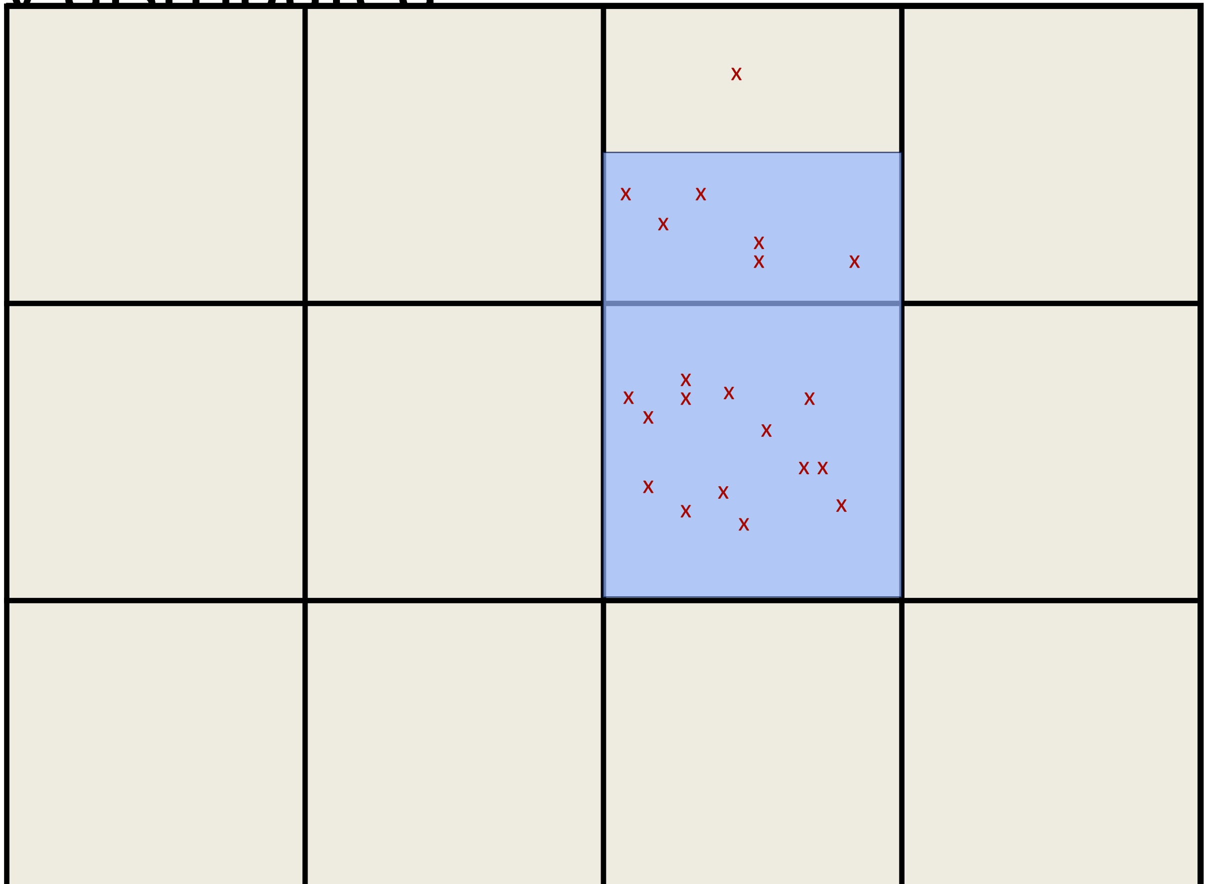
$$\sum^n \text{Var}\left(\text{Lap}(1/\epsilon)\right) = \frac{2n}{\epsilon^2}$$



Sources of Error

2. Error from grid: Non-uniformity error

- Assuming the data points within each cell are uniformly distributed



Error Minimization

- Noise error: calls for coarser partitioning
- Non-uniformity error: calls for finer partitioning
- Need to choose partition granularity to minimize the sum of the two errors

Determining Grid Size

- m x m grid. Query selects a portion r of the domain.
- Standard deviation of the noise error: $\frac{\sqrt{2rm^2}}{\epsilon}$
- Standard deviation non-uniformity error: $\frac{\sqrt{r}N}{c_0 m}$
- Minimize sum of two errors

$$\arg\min_{m} \frac{\sqrt{2r}m}{\epsilon} + \frac{\sqrt{r}N}{mc_0}$$

$$m = \sqrt{\frac{N\epsilon}{c}}, c \approx 10$$

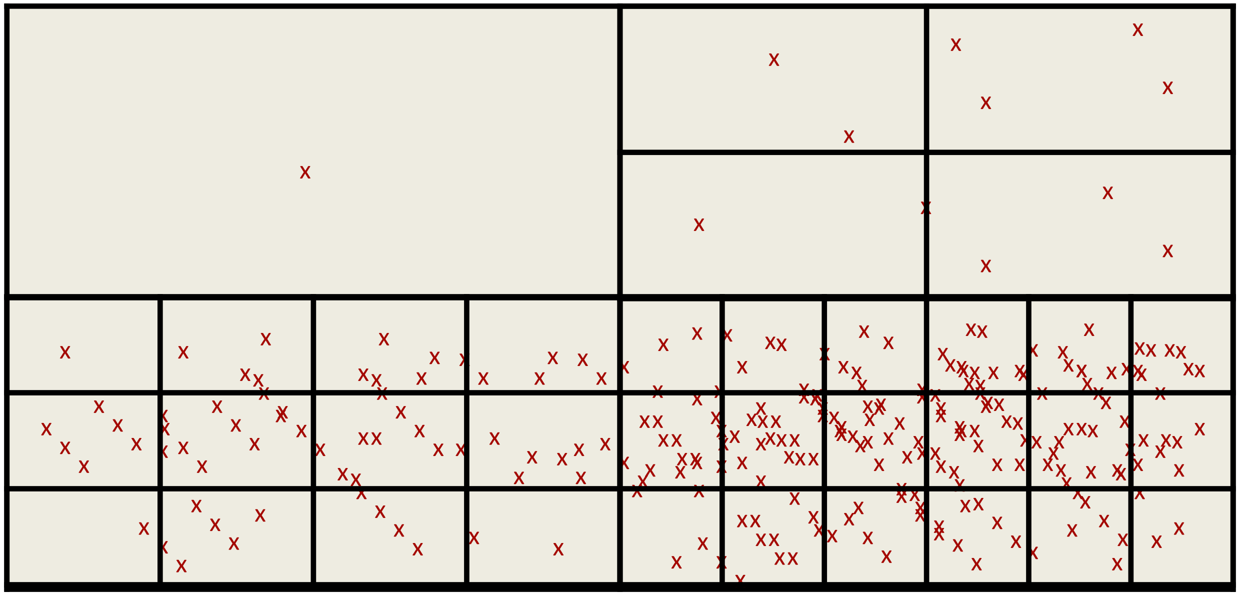
Limitation of Uniform Grid

- Uniform Grid treats all regions equally
  - If a region is *sparse*, we might *over*-partitioning the region. This increases the noise error with little reduction in the non-uniformity error.
  - if a region is very *dense*, this method might result in *under*-partitioning of the region. As a result, the non-uniformity error would be quite large.

Adaptive Grid

- Adapt the level of partitioning based on the number of data points in each region
  - If a region is dense, use finer granularity to reduce non-uniformity error
  - If a region is sparse, use a more coarse grid

Adaptive Grid



Adaptive Grid

- Two level partitioning:

$\alpha\epsilon$

1. Lay a coarse *m1* x *m1* grid over the data domain and obtain a noisy count for each cell

$(1-\alpha)\epsilon$

2. Partition each cell into an *m2* x *m2* grid, where *m2* depends on the noisy count of the cell
3. Apply constrained inference

Adaptive Grids

- Choosing Parameters (*m2*):
  - Average noise error:

$$\sqrt{\frac{(m_2)^2}{4}}\frac{\sqrt{2}}{(1-\alpha)\epsilon}$$

  - Average non-uniformity error: $\frac{N'}{c_0 m_2}$

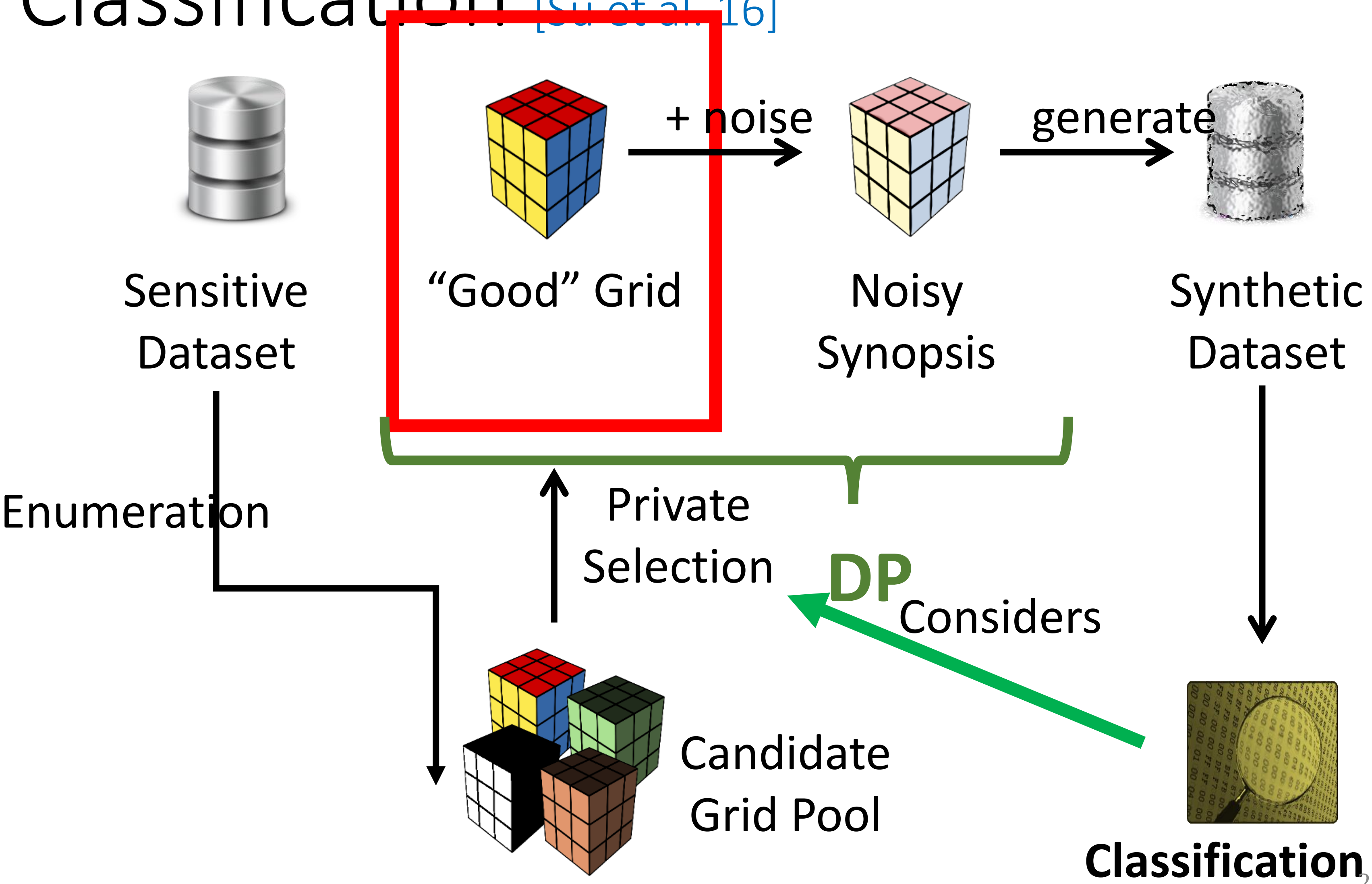
$$m_2 = \left\lceil \sqrt{\frac{N'(1-\alpha)\epsilon}{c_2}} \right\rceil$$

Adaptive Grids

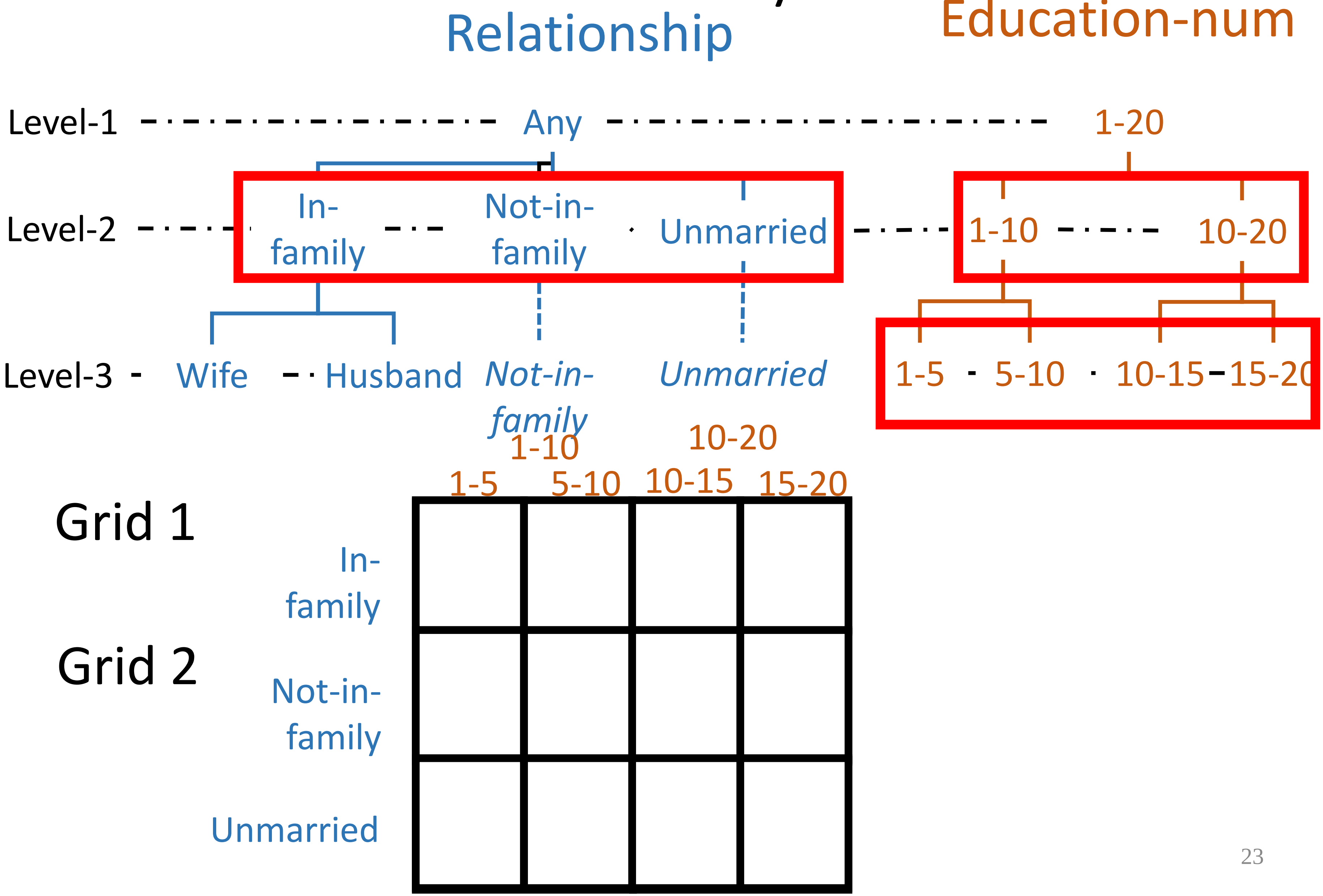
- Choosing Parameters ($m1$):
  - Parameter is less critical, since the second level adapts to the count of each cell
  - In general, we want it to be less than the choice for uniform grids.

$$m_1 = \max\left(10, \frac{1}{4}\left\lceil\sqrt{\frac{N\epsilon}{c}}\right\rceil\right).$$

PrivPfC: Private Publication for Classification [Su et al. 16]



Generation Hierarchy & Grids



Histogram Classifier

Dataset

Project

Grid 1

Education-num

Relationship	1-10	10-20
In-family	20|16	2|6
Not-in-family	7|5	2|8
Unmarried	6|2	9|3

Majority Voting

Consider the Noise Impact

<table>
<tr><th></th><th colspan="2">Education-num</th></tr>
<tr><th>Relationship</th><th>1-10</th><th>10-20</th></tr>
<tr><td>In-family</td><td>20|16</td><td>2|6</td></tr>
<tr><td>Not-in-family</td><td>7|5</td><td>2|8</td></tr>
<tr><td>Unmarried</td><td>6|1</td><td>9|3</td></tr>
</table>

+Noise

<table>
<tr><th colspan="2">Education-num</th></tr>
<tr><th>1-10</th><th>10-20</th></tr>
<tr><td>21|15</td><td>2|8</td></tr>
<tr><td>6|7</td><td>1|6</td></tr>
<tr><td>8|1</td><td>10|4</td></tr>
</table>

Label is flipped

Quality Function

- Number of correctly classified points is a random variable
- Use expectation as the quality

$$\mathrm{qual}(g) = \sum_{c \in g} n_c^+ \cdot p_c^+ + n_c^- \cdot (1 - p_c^+)$$

- n_c^+ : number of points in *c* with positive label
- p_c^+ : prob of positive label is the majority after injecting the noise
- $p_c^+ = \Pr[n_c^+ + Z_c^+ > n_c^- + Z_c^-]$ $Z_c^+, Z_c^- \sim \mathrm{Lap}(1/\epsilon)$,

ϵ is budget for perturbation

Quality Scores under Different Settings

Relationship \ Education-num	1-10	10-20
In-family	20 | 16	2 | 6
Not-in-family	7 | 5	2 | 8
Unmarried	6 | 1	9 | 3

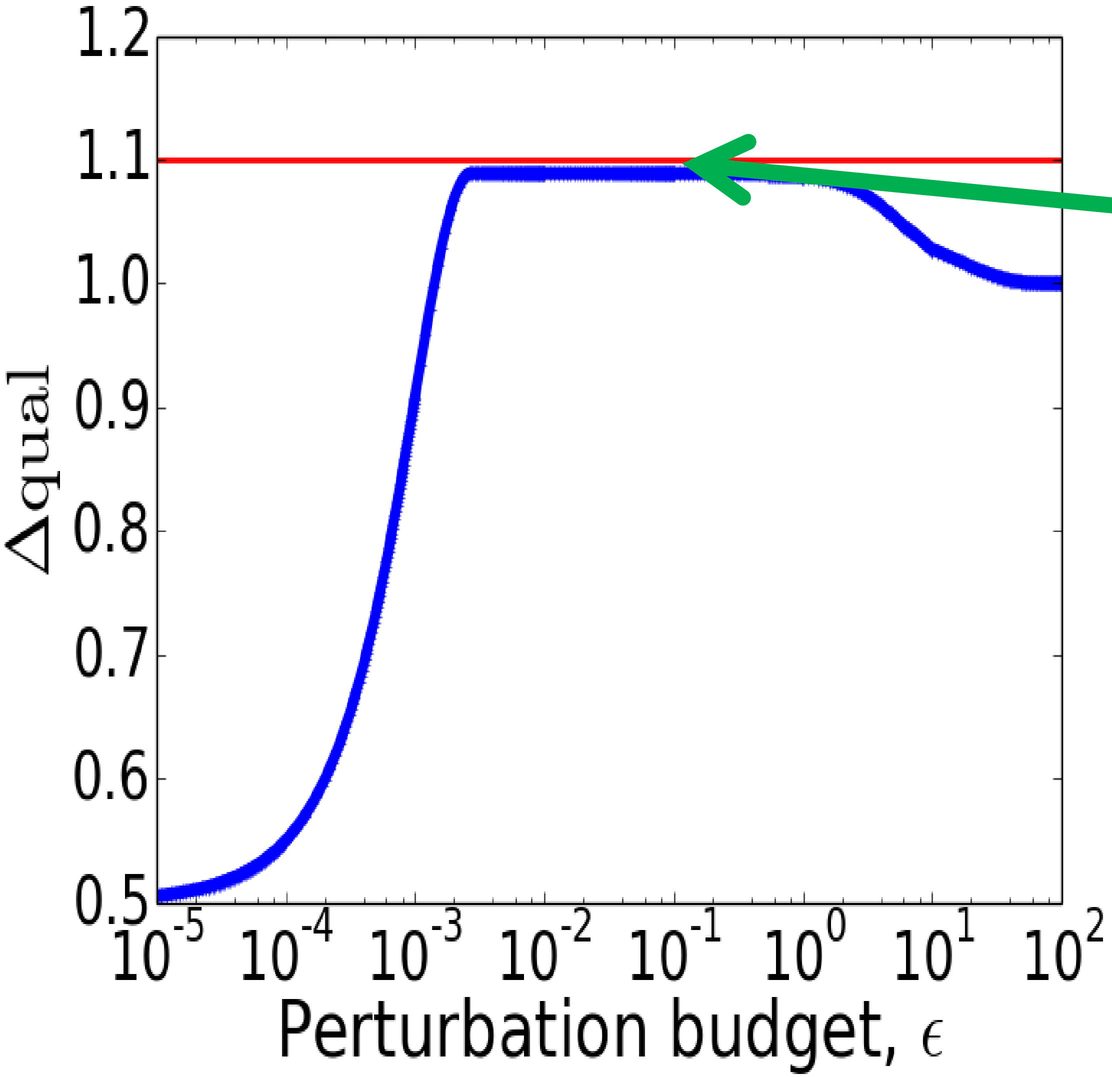
Relationship \ Education-num	1-5	5-10	10-15	15-20
In-family	14 | 8	6 | 8	2 | 0	0 | 4
Not-in-family	5 | 2	2 | 3	2 | 0	0 | 8
Unmarried	6 | 1	0 | 1	4 | 2	3 | 1

$\epsilon = 0.1$ / $\epsilon = 1.2$

$qual(g_1) = 45$ / $qual(g_1) = 57$

$qual(g_2) = 48$ / $qual(g_2) = 55$

Sensitivity of Quality Function



Sensitivity < 1.1

PrivPfC: Private Publication for Classification

Algorithm 7 PrivPfC: Differentially Privately Publishing Data for Classification

Input: dataset D, the set of predictor variables A and their taxonomy hierarchies, total privacy budget ϵ, maximum grid pool size Ω.

$\epsilon_N \leftarrow 0.03\epsilon$, $\epsilon_{sh} \leftarrow 0.37\epsilon$, $\epsilon_{ph} \leftarrow 0.6\epsilon$
$\hat{N} \leftarrow |D| + \mathrm{Lap}(1/\epsilon_N)$
$T \leftarrow 20\% \cdot \hat{N} \cdot \epsilon_{ph}$
$\mathsf{H} \leftarrow \mathrm{Enumerate}(\mathsf{A}, \Omega, T)$
Comment: privately select grid
for $i = 1 \rightarrow |\mathsf{H}|$ **do**
  $q_i \leftarrow \mathrm{qual}(\mathsf{H}_i)$
  $p_i \leftarrow e^{-(q_i \epsilon_{sh})/2}$
end for
$h \leftarrow$ sample $i \in [1..|\mathsf{H}|]$ according to p_i
Initialize I to empty Comment: privately perturb grid
for each cell $c \in h$ **do**
  $\hat{n}_c^+ \leftarrow n_c^+ + \mathrm{Lap}(1/\epsilon_{ph})$
  $\hat{n}_c^- \leftarrow n_c^- + \mathrm{Lap}(1/\epsilon_{ph})$
  Add $(\hat{n}_c^+, \hat{n}_c^-)$ to I
end for
Round all counts of I to their nearest non-negative integers.
return $\hat{I}$

Summary

	Differentially Private Optimization for ML	
	Single Workload	Non-interactive
Idea	Breaking task into query workload	Publishing a synopsis for answering any query
Pros	Customized for a specific task	- Re-usable - Preserve data distribution
Cons	- Over-divided privacy budget - No more data distribution	- Not customized for a specific task

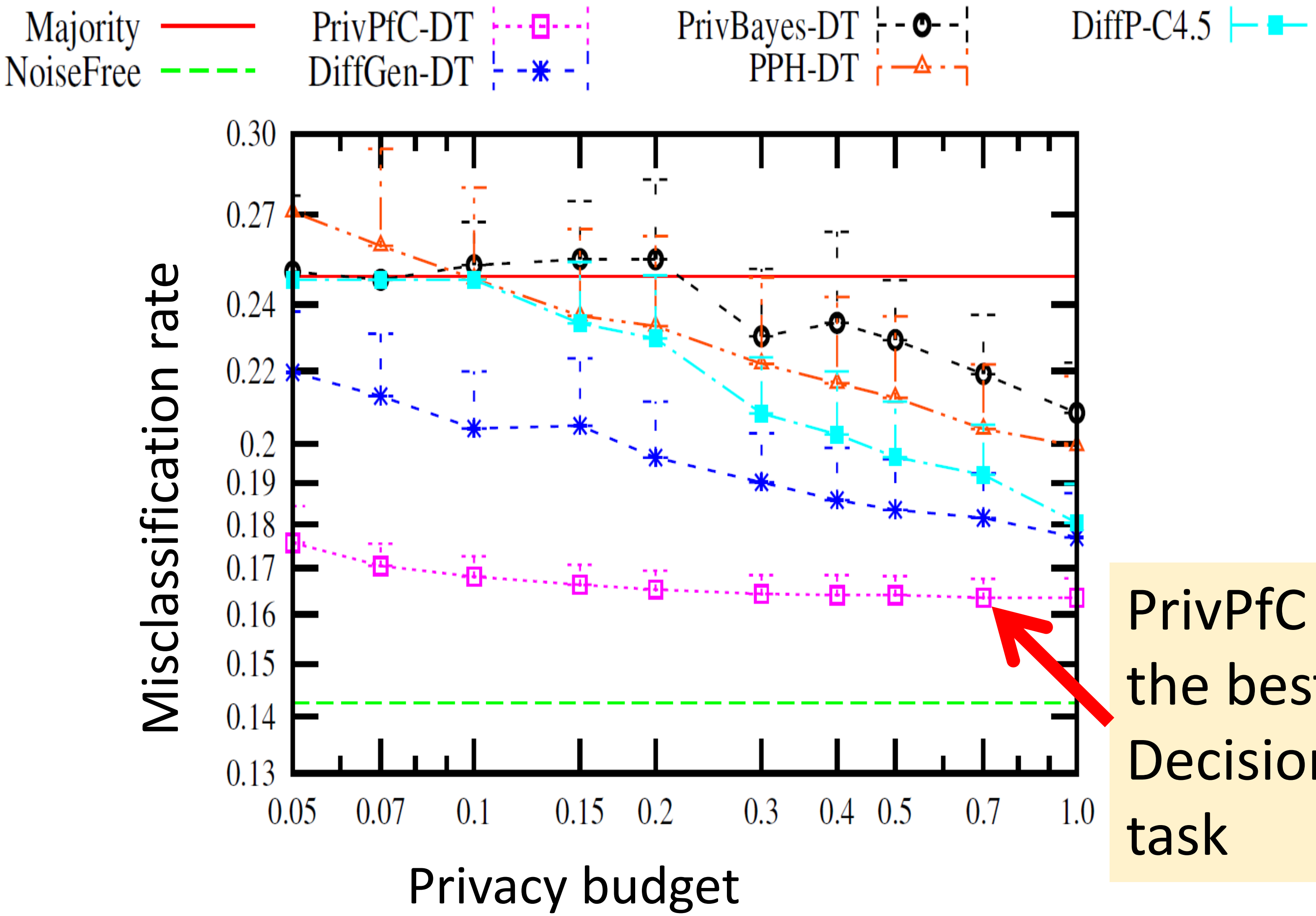
Experiments on DP Classification

- Datasets

Dataset	#Dim	#Num/#Cate	#Records	Task
Adult	15	6/8	45,222	>50K/yr
Bank	21	10/10	41,188	Sub. Deposit
US	47	15/31	39,186	>50K/yr
BR	43	14/28	57,333	>300/mo

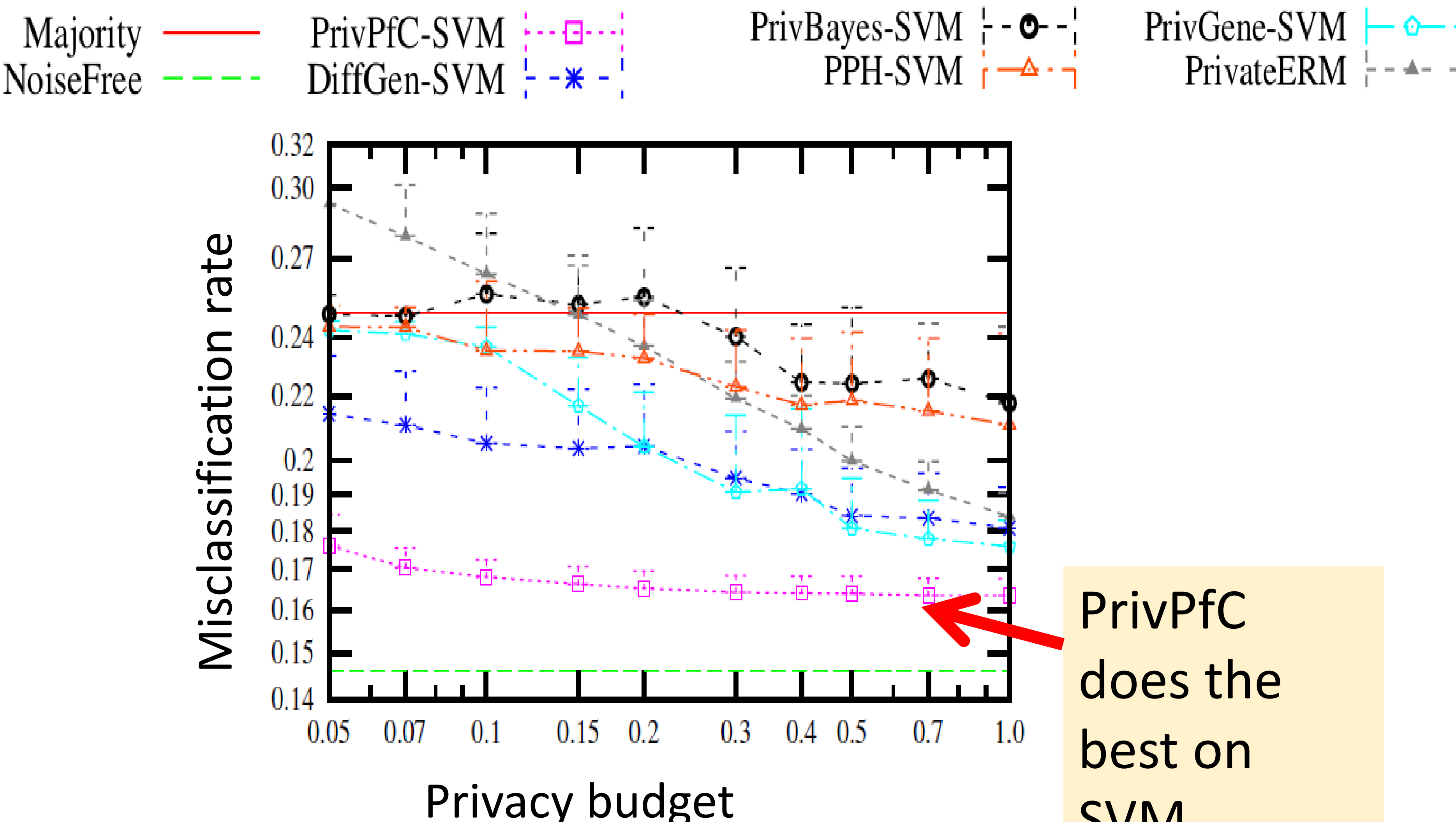
- Setup
  - Decision tree and SVM
  - Vary epsilon from 0.05 to 1.0
  - Misclassification rate

Decision Tree Comparison on Adult Dataset



PrivPfC does the best on Decision Tree task

SVM Comparison on Adult Dataset



PrivPfC does the best on SVM

Next Lecture

- Meanings and caveats of DP